Bases, Corpus et Langage - UMR 6039

CORPUS 1

2002

Corpus et recherches linguistiques

Sylvie MELLET

J. Guilhaumou

D. Mayaffre

A. Jaubert

J.-Ph. Dalbera

M. Plénat *et alii*

Les corpus *réflexifs* : entre architextualité et hypertextualité

Damon MAYAFFRE
« Bases, Corpus et Langage », UMR 6039, CNRS

Résumé : Un des enjeux actuels du traitement sémantique des corpus textuels concerne la nécessaire tentative de contrôle et d'objectivation de l'intertexte. Les *corpus réflexifs*, que nous définissons dans cet article, poursuivent cette exigence d'objectivation et de mise en forme des ressources sémantiques et interprétatives, en se proposant d'être, dans la mesure du possible, des tout-textuels sémantiquement auto-suffisants – c'est-à-dire des univers interprétatifs clos, définis parmi d'autres – pour une exploitation certes pas exhaustive, mais raisonnable et raisonnée du texte.

Cette contribution cherche à participer à l'élaboration de critères pertinents à la construction des grands corpus textuels électroniques dans les Sciences du langage, les Lettres et les Sciences humaines. Plus précisément, dans le cadre d'analyses pragmatiques des textes (analyses du discours et analyses textuelles, praxématique, commentaires littéraires, analyses de contenu, ...) et celui des traitements automatiques ou semi-automatiques des macro-corpus (TALN, lexicométrie), notre réflexion pointe – en espérant à terme le réduire – le grand écart épistémologique qu'effectuent les chercheurs entre le moment de l'établissement et du traitement du corpus, sur la base de critères explicites et objectifs, et le moment de l'interprétation du texte qui fait appel, souvent sans garde fou et sans autre précaution, à un intertexte subjectif et discrétionnaire au contour indéterminé.

L'on se propose en fait ici, à partir d'une définition consensuelle du corpus, de rappeler deux dimensions essentielles de l'objet (sa composition sérielle et sa nature heuristique) pour mieux souligner la limite scientifique que représente, dans nos pratiques interprétatives, l'articulation douloureuse entre le texte et

l'intertexte, le corpus et le hors-corpus. Ensuite, l'on avancera un critère nouveau de pertinence pour les corpus – la *réflexivité* ou l'hypertextualité sémantique – susceptible, dans une certaine mesure, de dépasser cette limite en proposant un nouveau parcours de lecture dans lequel l'acte interprétatif final doit être pressenti dans l'acte originel de la constitution même du corpus.

1. Définition

Qu'est ce qu'un corpus textuel ? A en croire la littérature spécialisée[1], à en croire tout simplement l'*Encyclopaedia Universalis* ou le *Robert*, un corpus textuel peut être défini simplement comme :

> « *Un rassemblement de textes ou une collection de textes regroupés sur la base d'hypothèses de travail en vue de les interroger.* »

De cette définition découlent deux idées simples mais fortes, évidentes mais essentielles qui poussées jusqu'au bout de leur logique débouchent non seulement sur la problématique de notre réflexion (le dépassement de la dialectique texte / intertexte) mais sur les bases des pratiques scientifiques pour l'ensemble des disciplines qui ont affaire aux textes dans une préoccupation herméneutique.

---

[1] On se reportera notamment à M.-P. Péry-Woodley (1995) « Quels corpus pour quels traitements automatiques ? », *Traitement Automatique des Langues*, n°36 (1-2), pp. 213-232 ; B. Habert *et al.* (1997), *Les linguistiques de corpus*, Paris : A. Colin ; A. Condamines *et al.* (1999), *Corpus et traitement automatique des Langues : pour une réflexion méthodologique*, Actes de l'atelier thématique TALN, Cargèse ; M. Bilger (éd.) (2000), *Corpus. Méthodologie et applications linguistiques*, Paris : Champion ; et en contre-point *Cahiers de praxématique* (1999), n°33, « Sémantique de l'intertexte ».

En tant qu'historien on citera évidemment l'ouvrage pionnier de R. Robin (1973), *Histoire et Linguistique*, Paris : A. Colin ; A. Prost (1988), « Les mots » *in Pour une histoire politique* (sous la dir. de R. Rémond), Paris : Seuil ; D. Peschanski (1984), *Histoire moderne et contemporaine et informatique*, n°4 spécial « Le corpus » ; J. Guilhaumou, D. Maldidier & R. Robin (1994), *Discours et archive*, Paris : Mardaga, 1994.

1.1. Perspective sérielle

D'abord, si l'on s'attache à la première partie de la définition, il faut admettre qu'à partir du moment où il y a « rassemblement » ou « collection » de textes, le chercheur, consciemment ou non, entre dans une perspective sérielle.

On additionne des textes, des données, et même si cette addition aboutit à un résultat qui a son unité et sa cohérence, la pluralité sérielle du corpus va dominer l'analyse. L'aspect composite, additionnel, sériel du corpus va devenir central dans la manipulation de l'objet. Quand bien même un philosophe ne rassemblerait que deux aphorismes de *Par delà le bien et le mal*, les questions qu'il se posera devant ce corpus seront immédiatement du type : quels sont les points communs entre les deux textes ? Y a-t-il entre eux, rupture ou continuité, redondance ou complément, divergence ou convergence ? Que se passe-t-il lorsqu'on croise ces deux unités textuelles ? etc. En effet, si le chercheur ne s'interrogerait pas de la sorte, pourquoi aurait-il rassemblé ces textes dans son corpus de recherche ? Et quand bien même il nous paraîtrait faire l'économie de ces questions au cours de l'analyse, elles auront néanmoins présidé à la constitution du corpus.

Donc le corpus, parce qu'il est une composition – au sens d'addition –, nous fait entrer, qu'on le veuille ou non, dans une perspective sérielle. Or la notion même de série soulève au moins deux problèmes.

Premièrement elle pose cruellement la question de la méthode ou même de la technique de traitement. Si dans toute étude scientifique la dissection de l'objet est évidemment problématique, lorsque cet objet est composite le phénomène apparaît plus crûment. Comme en chimie, étudier ou disséquer un corps complexe nécessite infiniment plus de ressources scientifiques qu'étudier un corps simple. Or le corpus est – par définition – un corps complexe[2].

Il n'est pas le lieu ici de s'étendre sur cette première difficulté puisque les méthodes de traitement des corpus ne sont pas

[2] Nous laissons ici de côté le cas critique d'un corpus comptant un seul texte, d'un corpus représentant une série d'un seul élément – un seul mot, pourquoi pas ? – ou pire encore une série de 0 élément : un corpus vide

l'objet de cet article, mais ceux qui connaissent les travaux de l'UMR 6039 « *Bases, Corpus et Langage* », devinent les implications sous-jacentes à une telle démonstration : dans une perspective sérielle, si le corpus est une série de textes – et pour peu que celle-ci soit un peu importante –, l'approche quantitative apparaît difficilement contournable. A titre personnel, lorsqu'il s'est agi d'embrasser le vocabulaire de discours politiques d'une série de 832 discours représentant une cinquantaine de livres de poche, il a paru nécessaire d'avoir recours à la statistique lexicale et à la lexicométrie[3].

Deuxièmement le corpus en tant que série pose, en amont du problème du traitement, la question de sa constitution et plus précisément de sa *clôture*. C'est sur ce deuxième point qu'il faut ici s'arrêter.

Où commence et où s'arrête une série ? Ou plutôt, comment débuter et arrêter une série ? L'addition des deux aphorismes pressentis plus haut est-elle légitime et simplement efficace ? Plus grave, revêt-elle un caractère représentatif (de l'auteur, du genre, d'une époque ...) qui en justifierait l'étude ou lui donnerait du volume ? Ne faudrait-il pas lui ajouter un troisième aphorisme ou dix de plus, et ajouter encore toute la production textuelle de Nietzsche pour que ce corpus prenne de la pertinence ?

Devant la gravité de ces interrogations, le recours à l'échantillonnage est apparu longtemps comme une solution : une série limitée de textes (ou de bouts de textes) pouvait prétendre représenter un tout plus important ; l'en-dedans limité du corpus témoignerait d'un en-dehors plus vaste. En fait, à l'usage, l'échantillonnage constitue souvent un trompe-l'œil statistique insatisfaisant dans la plupart des Sciences humaines lorsqu'il s'agit de réellement interpréter les textes, de passer du texte au monde et de rendre compte des intentions pragmatiques du locuteur. De manière plus générale, la linguistique de corpus en soulignant la vanité des corpus échantillonnés prétendant être représentatifs de la Langue ou la linguistique textuelle en insistant sur l'infinie

[3] D. Mayaffre (2000), *Le poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres. Maurice Thorez, Léon Blum, Pierre-Etienne Flandin et André Tardieu (1928-1939)*, Paris : Honoré Champion. On lira dans l'introduction, IV) Corpus, pp. 38-51.

variation (générique, stylistique, politique ...) de la Parole mettent à jour *l'auto-insuffisance* des corpus, en leur clôture, dans le cadre de la sémantique du texte, au moment de l'acte interprétatif[4].

Ainsi, additionner en matière de corpus signifie de manière problématique avant tout soustraire. Décider de rassembler deux textes, c'est avant tout décider d'écarter tous les autres. Or, répétons-le, comment juger que le rassemblement établi est non seulement nécessaire (ou utile) mais suffisant ?

Et d'abord, nécessaire et suffisant dans quel but ? Pour quoi faire ?

1.2. Un objet heuristique

Cette dernière interrogation nous amène à la deuxième idée-force que porte en elle la définition du corpus comme une collection de textes regroupés sur la base d'hypothèses de travail en vue de les interroger.

L'aspect redondant de la deuxième partie de la définition insiste sur un point. Le corpus est un objet heuristique. C'est une construction arbitraire, une composition relative qui n'a de sens, de valeur et de pertinence qu'au regard des questions qu'on va lui poser, des réponses que l'on cherche, des résultats que l'on va trouver.

Ce n'est pas un donné disciplinaire mais un objet heuristique. Le contenu objectif ou matériel d'un corpus textuel n'appartient pas à l'Histoire, à la Linguistique ou à la Philosophie. C'est l'intention du chercheur qui est importante et lui donne son sens.

Peut-être peut-on penser que le corpus de *Par delà le bien et le mal* est un corpus objectivement philosophique ? Evidemment, il n'en est rien. Un historien aura pu rassembler ces textes pour chercher, par exemple – vieille rengaine – si la dénonciation de la morale judéo-chrétienne chez Nietzsche a, oui ou non, fait le lit de l'idéologie nazie en Allemagne. Un linguiste pourra les avoir

---

[4] B. Pincemin (1999), « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative », *in* A. Condamines *et al.*, *Corpus et traitement ..., op. cit.* ; J.-M. Adam (1999), *Linguistique textuelle. Des genres de discours aux textes*, Chap. 3 : *Types de textes ou genres de discours ?* Paris : Nathan, pp. 81-100.

rassemblés pour étudier le fonctionnement des marques énonciatives ou de l'utilisation de l'imparfait, etc.

Evidemment poser le corpus comme objet heuristique ouvre des problématiques épistémologiques importantes qui dépassent les limites de nos compétences. L'heuristique en tant que quasi-discipline renvoie à l'histoire de la philosophie elle-même et à la plupart des grands auteurs d'Aristote à Ricœur, en passant par Bacon, Popper ou Bachelard. Face à cela, il convient avant tout de rappeler la modestie de notre position.

Néanmoins si l'on peut schématiser ce que l'heuristique – cette théorie et cette pratique de la recherche et de la découverte scientifiques – implique au niveau du corpus, on peut retrouver les deux éléments abordés précédemment à propos du corpus comme composition sérielle : la question méthodologique et le problème de la clôture ou de l'auto-(in)suffisance des ressources du corpus au moment de l'acte interprétatif.

L'heuristique s'interroge avant tout, dès l'origine, sur les techniques pour extraire des résultats. C'est, semble-t-il, la préoccupation première d'Aristote dans l'*Organon*. La découverte passe par des *techniques* heuristiques, par un *ars inveniendi*, à terme, dans l'évolution des sciences, par une *méthode* constituée de traitement. Il faut un protocole méthodologique – une procédure intellectuelle et des procédés techniques explicites – pour traiter un corpus.

Ceci mérite évidemment d'être rappelé en ce qui concerne les corpus textuels car le texte ou de manière plus générale le langage paraît à ce point inné et naturel que les chercheurs – en Histoire par exemple[5] – ont tôt fait de le traiter, sans médiation scientifique, sans autre forme de procès : intuitivement.

Il faut à nouveau laisser de côté la réflexion méthodologique. Précisons cependant que si la constitution du corpus sur la base d'hypothèses de travail et selon des critères bien définis (homogénéité, contrastivité ...) est le premier moment d'une recherche scientifique, la mise en place d'une procédure de

[5] D. Mayaffre (1999), « Histoire politique ou politique de l'Histoire ? », *La Pensée*, n°319, pp. 139-147.

traitement en est le second. On peut même dire que ce sont deux phases essentielles, concomitantes et réversibles ; en tout cas intimement liées. Le corpus commande la méthode et la méthode ordonne le corpus. Si l'on dispose d'un petit corps de texte, on exclura l'approche statistique pour se tourner vers des méthodes qualitatives ; si l'on est le tenant de la méthode statistique, on établira un corpus adéquat sur lequel cette approche sera opérationnelle.

Mais l'heuristique pose aussi, à un niveau supérieur, le problème crucial de l'interprétation. L'heuristique générale prise dans le sens baconien peut même être définie comme une tentative de compréhension et d'interprétation du monde (ou ici, prosaïquement, d'une compréhension ou d'une interprétation du corpus) : le monde tout entier (ici le corpus) est à interpréter ; la recherche scientifique a pour fonction de lui donner sens.

Or le moment de l'interprétation du corpus – moment qui constitue bien le nœud gordien des études textuelles dans les sciences molles – nous renvoie subrepticement à nouveau à la question de sa clôture ; et plus précisément au grincement de l'articulation entre l'intérieur et l'extérieur du corpus ; à la tension dialectique – pour ne pas dire contradictoire – entre la prise en compte (le traitement) des ressources intérieures du corpus et l'appel à des ressources qui lui sont extérieures pour le comprendre et l'interpréter.

Soit : nous avons un corpus ; soit encore : nous avons échafaudé une méthode pour l'analyser ; mais est-ce suffisant pour interpréter ses textes ? N'a-t-on pas besoin d'éléments étrangers – disons pour l'instant de la manière la plus vaste et la plus problématique, de notre culture – pour le comprendre ? Et ainsi ne se trouve-t-on pas projeté en dehors des limites objectives du corpus, dans un tout-subjectif, sans garde fou scientifique ni garantie méthodologique ?

Ces questions sont impérieuses à propos d'un corpus textuel car les ressources extérieures que l'on va tout à coup mobiliser pour interpréter le texte du corpus seront elles-mêmes le plus souvent du texte. Elle seront donc de même nature que le corpus, entraînant une confusion et une discrimination difficilement

justifiable entre les textes du corpus traités scientifiquement et ceux de l'intertexte mobilisés sans traitement préalable.

Pour être caricatural, on peut reprendre, une dernière fois, l'exemple qui nous poursuit. On juge suffisant deux textes de Nietzsche, on les rassemble sur la base d'hypothèses de travail explicites et selon des critères clairement définis (même longueur par exemple, même registre, même auteur...), on les analyse selon une méthode établie afin d'en faire une exégèse scientifique, on en extrait des résultats bruts aussi objectifs que la méthode le garantit (une liste des fréquences par exemple dans le cadre d'un traitement lexicométrique), mais voilà qu'au moment de les interpréter, on fait appel et référence à tels textes de Spinoza ou de Hegel, on convoque, en complément, d'autres ouvrages de Nietzsche. Autant de nouveaux textes pour lesquels on n'aura pas fait l'effort de définition, de rassemblement, de traitements scientifiques pour les manipuler, les analyser et, eux-mêmes, les comprendre.

La linguistique de corpus ou la sémantique du discours ont définitivement démontré que pas plus qu'on ne pouvait comprendre un mot sans la phrase et la phrase sans le discours, on ne pouvait comprendre le discours sans l'interdiscours, le texte sans le co-texte (sans même parler ici du hors-texte[6]), c'est-à-dire aussi et de manière plus générale, le corpus sans le hors corpus[7].

Tant est si bien que les efforts scientifiques qui président à la constitution des corpus textuels comme à leurs traitements rigoureux semblent anéantis au moment du bond interprétatif qui nous projette dans la lave d'un intertexte indéterminé, appréhendé intuitivement.

---

[6] Nous ne parlerons pas en effet ici du problème insoluble pour les linguistes de la prise en considération pragmatique du contexte socio-historique des textes. Nous nous bornerons au problème suffisamment difficile de la considération du contexte linguistique autrement appelé co-texte.

[7] On lira à ce sujet, J. Habermas (1994), *Textes et contextes*, Paris : Seuil ; J. Caron (1995), « Signification, interprétation et contexte », *in Les rôles du contexte et de la situation dans la cognition*, Doc. de la 5ème école d'été de l'ARC, pp. 30-32 ; F. Rastier (1998), « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage. », *Langages*, n°129, pp. 97-111 ; et du même auteur, *Arts et sciences du texte* (2001), Paris : PUF, notamment le chapitre III « Philologie numérique », pp. 91-99.

Nos pratiques de sémantique interprétative semblent ainsi comprises dans un cercle carré où le chercheur se trouve déchiré entre le cocon objectif que constitue l'en-dedans du corpus, minutieusement rassemblé et analysé, et le tout-subjetif de son en-dehors textuel nécessaire à son interprétation.

2. Proposition

Précisément, le deuxième temps de cet article veut essayer de proposer un critère nouveau de pertinence des grands corpus électroniques susceptible de réduire ce grand écart épistémologique entre le texte du corpus soumis à une méthode et l'intertexte mobilisé et appréhendé intuitivement.

Les réflexions sur la constitution des corpus se trouvent aujourd'hui profondément modifiées par l'apparition de macro-corpus. Le développement des performances des ordinateurs et des logiciels d'une part, la disponibilité effective de bases de données textuelles toujours plus vastes d'autre part, permettent aux chercheurs désormais d'envisager de traiter des corpus de plusieurs dizaines de millions de mots. Ce bond quantitatif, vivifiant en lui-même, entraîne une révolution qualitative dont on doit tirer profit.

2.1. Constituer de grands corpus réflexifs

Au-delà des critères de pertinence habituels – homogénéité, contrastivité et diachronicité – sur lesquels nous ne pouvons revenir[8] mais que l'on peut revisiter à la lumière des grands corpus[9], un corpus aujourd'hui, par sa taille et la compilation souvent exhaustive des textes afférents à un domaine donné, gagne à être *réflexif*.

Nous entendons par *réflexivité* du corpus le fait que ses constituants (articles de presse, discours politiques, pièces de théâtre ; de manière plus générale, sous-parties) renvoient les uns aux autres pour former un *réseau sémantique* performant dans un tout (le corpus) cohérent et auto-suffisant.

---

[8] Cf. D. Mayaffre (2001), *Le poids des mots, op. cit.*, pp. 38-42.

[9] Il est évident par exemple que l'exigence d'*homogénéité* des textes du corpus demande à être redéfinie (assouplie) dans le cadre de corpus comptant plusieurs milliers de textes et plusieurs dizaines de millions de mots.

Cette notion de *corpus réflexif* pressentie dans notre thèse a été développée grâce à l'intervention de J. Guilhaumou lors de la soutenance. Le fait que, dans le même temps, S. Lahlou et H. Floch de l'ENS Fontenay-Saint-Cloud ont ébauché, en des termes certes différents, le même type d'analyse[10] ou que l'équipe de G. Vignaux réfléchisse à une notion, plus technique mais voisine, que constitue l'hypertextualité[11], et encore que L. Tanguy et Th. Thlivitis, avec le concept d'anagnose, en appellent à une formalisation de l'intertexte[12] laisse espérer qu'une piste de travail est désormais ouverte en direction d'une tentative d'objectivation de l'intertexte et d'organisation de parcours contrôlés de lecture (en ce qui nous concerne, auto-centrés sur le corpus) dans lesquels l'acte interprétatif est rendu si ce n'est objectif en tout cas transparent.

Le caractère réflexif des grands corpus textuels apparaîtra, selon nous, comme une qualité déterminante tant du point de vue de leur traitement purement linguistique que de leur traitement socio-linguistique.

Sans même ici parler, de manière trop ambitieuse pour la proposition avancée, de toutes les ressources textuelles interprétatives nécessaires à l'interprétation d'un texte – *l'intertexte* au sens le plus général et le plus élargi –, il est donc admis d'un point de vue linguistique que le sens des mots et des discours ne peut être saisi sans une remise en contexte grâce à une étude des conditions d'énonciation. Dans ces conditions d'énonciation, l'environnement linguistique « immédiat » du texte – *le co-texte* – joue un rôle toujours central. Dans le champ clos du discours politique par exemple, une production verbale est prise de manière évidente dans une toile de discours déjà énoncés, de phrases déjà prononcées, de

---

[10] S. Lahlou & H. Folch (1998), « Quelques stratégies pour l'exploitation en ADT de grands corpus hétérogènes » *in JADT 1998*, Nice : CNRS-UNSA, pp. 381-391.

[11] Laboratoire *Communication et Politique* (UPR 36 - CNRS) dirigé par Georges Vignaux ; équipe « Hypertextes et textualité électronique ».

[12] T. Thlivitis (1999), *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension de textes*, Thèse de Doctorat Informatique, Univ. De Rennes 1 ; et L. Tanguy & T. Thlivitis (1999), « Parcours interprétatifs (inter)textuels : vers une assistance informatique », *Cahiers de praxématique*, n°33, pp.185-215.

mots déjà utilisés par rapport auxquels cette production se situe et auxquels elle répond parfois directement[13].

Précisément, c'est ce co-texte qui doit, autant que faire ce peut, désormais se trouver intégré dans le corpus lui-même. Ou, autrement dit, les macro-corpus en embrassant la plupart des discours ou textes d'un sujet donné, d'un locuteur donné, d'une période donnée compteront automatiquement le co-texte des textes qui le composent : le co-texte des textes du corpus sera le corpus.

L'avantage est évident. Il ne sera plus nécessaire de sortir du corpus pour comprendre et interpréter ses composants. Et l'analyse contextualisée ou co-textualisée de chacun des textes se fera grâce à une navigation interne au corpus et non sur la base de ressources extérieures arbitrairement et subitement convoquées.

Du point de vue de l'épistémè des recherches en Sciences humaines (Sociologie, Histoire...) les conséquences attendues ne sont pas moins grandes. Par la constitution de macro-corpus réflexifs, corpus et archive pourront se confondre en grande partie[14]. L'objet d'étude (tel discours politique par exemple) et les sources (les archives éclairant ce discours, constituées d'autres discours, d'autres prises de parole, et en Histoire toujours, par définition même de l'archive, d'autres productions textuelles) seront rassemblés et reliés dans le corpus, traités d'un même mouvement, par une même méthode. L'historien là encore n'aura pas ou moins à sortir de son corpus et de son traitement scientifique pour l'éclairer par un travail spécifique d'archive. Et le travail même d'archive sera partie intégrante du travail de saisie et de constitution du corpus.

Plus largement encore, le développement de gros corpus réflexifs et enrichis permet d'envisager, à terme, l'intégration de la bibliographie dans le corpus lui-même. Un discours sera alors relié

---

[13] Sans même parler, comme le montre M. Bakhtine, des discours de l'adversaire que l'on pressent et que l'on veut parfois prévenir : « *Intentionnellement ou non, chaque discours entre en dialogue avec les discours antérieurs tenus sur le même objet, ainsi qu'avec les discours à venir, dont il pressent et prévient les réactions* » (T. Todorov (1981), *Mikhaïl Bakhtine. Le principe dialogique*, Paris : Seuil, p. 8).

[14] On lira ainsi sous un jour nouveau les réflexions sur le corpus dans l'ouvrage référence de J. Guilhaumou, D. Maldidier & R. Robin (1994), *Discours et archive*, *op. cit.*

non seulement aux textes qui constituent son co-texte, mais aux écrits scientifiques permettant de mieux le comprendre.

En attendant ce développement final, nous voulons insister sur l'aspect effectif de cette réflexion. Le corpus de notre thèse était un corpus qui tendait vers la réflexivité. L'importance des quatre locuteurs choisis sur l'échiquier politique faisait que les prises de parole de chacun renvoyaient souvent précisément aux prises de parole des trois autres ; souvent Thorez cite littéralement une intervention de Blum et les discours de Flandin reprennent parfois un article de Tardieu, pour former, à l'intérieur du corpus, un jeu de miroir linguistique intéressant. A une même date et parfois dans un même lieu (l'Assemblée nationale par exemple) nous nous sommes appliqué à saisir les discours des quatre personnages, qui s'articulaient comme un jeu de questions-réponses sémantiques, parfois explicite, souvent implicite, toujours avéré. Dans tous les cas, il a été intéressant de démêler cet écheveau linguistique en repérant les enjeux rhétoriques, les reprises lexicales volontaires et involontaires, les redéfinitions métalinguistiques ou les déformations sémantiques possibles. Au départ, le corpus avait simplement l'ambition de permettre des comparaisons froides et de faciliter l'étude de la partition de chacun au regard de celle des autres (corpus contrastif) ; en fait ce sont souvent les interactions linguistiques de cet ensemble polyphonique (corpus réflexif) qui ont retenu notre intérêt.

De même, du point de vue du travail historique d'archive, tel discours faisait référence à tel événement. Et pour juger de l'événement et de la manière dont il était traité par le locuteur, il convenait d'avoir d'autres sources. Comme souvent en histoire politique, les articles de journaux de l'époque paraissaient l'outil idéal et comme nous avions saisi beaucoup d'éditoriaux de Blum dans *Le Populaire*, de Thorez dans *l'Humanité*, de Tardieu dans *Gringoire* ou de Flandin dans le *Journal*, il a été facile de collecter des renseignements dans le corpus même pour reconstituer la trame événementielle ou contextuelle qui nous intéressait[15].

[15] C'est ici qui nous rejoignons S. Lahlou et H. Folch qui, pour étudier le thème du service public dans un macro-corpus, notent : « *Le premier problème est la remise en contexte historique ("hors texte") des textes que nous collectons.*

Tout ceci a été empirique et le traitement – la mise en relation des textes ou de traits linguistiques entre eux à l'intérieur du corpus – a été fait seulement de manière semi-automatique. Mais même artisanalement et à petite échelle l'intérêt nous a semblé évident.

Le 7 juin 1932, dans un lourd débat, à la Chambre, sur la politique étrangère de la France, André Tardieu déclare sans plus de précision :

> « *Entre la thèse que j'ai défendue à Genève, qui est celle depuis toujours de tous les gouvernements français, et la thèse de M. Léon Blum, il y a, je le répète, un abîme. Je demande donc simplement à M. Le président du Conseil* [Herriot] *de nous dire s'il accepte oui ou non la thèse de M. Léon Blum...* (silence). *Je prend acte de votre silence.* »[16]

Ce passage (caricatural à dessein pour les besoins de la démonstration) n'est compréhensible, en lui-même, ni dans le cadre de la linguistique pragmatique ni moins encore dans celui de l'histoire politique. L'analyse du sens profond de cet extrait réclame, ici de manière évidente, le recours à des ressources co-textuelles que l'on ne saurait envisager de traiter différemment que les ressources intratextuelles du passage : heureusement tant le discours de Blum dans lequel le dirigeant socialiste expose sa thèse que le discours à Genève devant la SDN de Tardieu où il exposait la

---

La compréhension d'un texte nécessite la connaissance du monde, ne serait-ce que parce que les objets auxquels il se réfère ne sont pas toujours explicités en détail dans le texte. C'est vrai pour une interprétation strictement linguistique, et l'est encore plus pour une interprétation sociologique ... A l'origine, nous voulions construire une chronologie des événements, les décrivant, séparée de la base textuelle et sur laquelle pointeraient les textes. Par exemple, la directive européenne sur le marché de l'électricité ... Puis nous avons réalisé que les descriptions d'événements étaient elles-mêmes des textes, voire des événements discursifs, comme la directive, ou des prises de positions publiques de tel ou tel acteur. ***Nous avons donc pris la décision de considérer la base comme une série de documents qui se renvoient les uns aux autres. Ceci fait de nos collections des hypertextes à plusieurs voix, dont certaines, pas toutes, sont celles de nos acteurs internes.*** » (*art. cit.*, pp. 385-386).

[16] A. Tardieu, *JO Chambre-débats*, 7 juin 1932.

sienne font partie de notre corpus de textes et ont été soumis eux-mêmes à l'analyse.

2.2. Réflexivité ou hypertextualité sémantique ?

D'un point de vue sémantique, les corpus réflexifs peuvent donc être envisagés, avec François Rastier, comme des *architextes* :

> « *A un palier encore supérieur* [i.e. : celui du corpus], *on peut formuler un principe d'architextualité : tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent* »[17].

Ceci implique des procédés méthodologiques et techniques de mise en forme qu'il convient de développer mais dont on pressent déjà bien l'orientation. Les corpus réflexifs devront être organisés techniquement comme des *hypertextes* : chaque texte constituant devra être relié aux textes considérés comme parents. L'encodage sous la norme SGML (Standard Generalized Markup Language) et ses applications HTML (Hyper Text Markup Language) ou XML (Extensible Markup Language) apparaissent *a priori* comme les plus simples et les plus universels pour créer ces liens hypertextuels sur de grosses bases de données, tout en permettant un traitement lexicométrique traditionnel à un niveau de granularité plus fin (habituellement le mot). A l'instar de la navigation sur Internet, la navigation dans le corpus pourra alors s'effectuer[18]. Le but étant, répétons-le, que chaque texte puisse être

---

[17] F. Rastier (2001), *Arts et sciences du texte*, Paris : PUF, p. 92.

[18] On lira à ce sujet S. Heiden (1999), « Encodage uniforme et normalisé de corpus. Application à l'étude d'un débat parlementaire », *Mots*, n°60. On notera aussi qu'un logiciel comme *Hyperbase* a ouvert la voie depuis plusieurs années aux études contextualisées des mots par le développement d'un hypertexte permettant une navigation entre l'index des formes du corpus et les textes du corpus. En ce qui nous concerne, on remarquera qu'à partir du repérage quantitatif d'un mot dans un texte (disons le sur-emploi de « fascisme » dans un discours de Thorez) on pourra convoquer systématiquement non seulement l'ensemble des contextes syntagmatiques de « fascisme » dans la prose de Thorez mais aussi, sans difficulté technique, dans l'ensemble du corpus c'est à-dire dans les discours de Blum ou des deux dirigeants de droite. Tant est si bien que l'on pourra prétendre faire une

replacé dans son réseau co-textuel utile et nécessaire pour sa compréhension et son analyse.

C'est par ces considérations techniques que l'on retrouve les travaux de l'équipe « Hypertextes et textualité électronique » du Laboratoire *Communication et Politique* (UPR 36 - CNRS), dirigée par Georges Vignaux, dont le but, sur le site expérimental Claude Bernard[19], est bien de fournir avec un texte (*L'introduction à l'étude de la médecine expérimentale*, 1865) un intertexte objectif et contrôlé, nécessaire à sa compréhension. L'hypertextualité apparaît ainsi comme la réalisation pratique ou la mise en forme de la réflexivité. Précisons néanmoins que si l'hypertexte du site Claude Bernard offre bien, *d'un point de vue technique*, une liberté (légitimement contrôlée) du parcours interprétatif, il s'organise, *d'un point de vue cognitif*, comme la juxtaposition centripète – donc orientée – de ressources extérieures permettant de nourrir la compréhension d'un texte-père qui est le centre virtuel avoué du site et sa raison d'être.

Nous insistons pour notre part sur la circularité fertilisante, sans entrave et sans direction pré-établie, des informations sémantiques d'un corpus réflexif, certes auto-centré sur lui-même mais non focalisé *a priori* sur un de ses constituants.

Le corpus ne sera alors plus seulement considéré comme un rassemblement inanimé de textes à disséquer sous la lumière crue de projecteurs extérieurs, mais comme un outil dynamique et interactif permettant de construire du sens et des connaissances grâce à ses ressources internes et à ses richesses intrinsèques. En d'autres termes, il ne représentera plus seulement la chambre froide pour la dissection du texte mais, pour reprendre les mots de L. Tanguy et Th. Thlivitis à propos de l'intertexte, comme *« un lieu privilégié de l'acte interprétatif lui-même »*[20].

---

analyse sémantique « réflexive » si ce n'est complète en tout cas approfondie de ce mot fort du discours politique.

La mise au point de cette lexicométrie contextualisée et, dans un corpus contrastif, « inter-dynamique » préfigure l'organisation de la réflexivité sémantique des corpus réclamée ici.

[19] http://lcp.damesme.cnrs.fr/claude-bernard/

[20] L. Tanguy & T. Thlivitis (1999), « Parcours interprétatifs (inter)textuels : vers une assistance informatique », *Cahiers de praxématique*, n°33, p. 185.

Conclusion

Il faut pour conclure faire un effort de lucidité sur la réflexion engagée ici autour de la clôture des corpus et sur la proposition avancée pour leur donner un nouveau contour.

Souligner, à la lumière de l'articulation problématique texte / intertexte ou de manière un peu plus réductrice texte / co-texte, la difficulté pratique d'interpréter le texte n'a rien d'original. Et la subjectivité inévitable qui a paru être dénoncée dans la sémantique interprétative est peut-être la force et la richesse des études du texte dans les disciplines qui ont affaire au langage.

D'autre part, le concept proposé n'a pas la prétention d'être une panacée. A vrai dire, il s'agit plus d'un idéal que d'un objectif atteignable. Ni l'environnement co-textuel nécessaire à la compréhension d'un texte, ni moins encore les ressources inter-textuelles nécessaires à son interprétation[21] ne peuvent se laisser enfermer dans un corpus aussi grand soit-il. Au niveau supérieur, l'intertexte est effectivement un objet dynamique, « *d'une nature opportuniste* »[22], en perpétuelle reconstruction selon l'état de la science et plus encore selon les objectifs du chercheur. Pour cette raison, il est non seulement difficile de l'enfermer dans un corps de textes définitivement établi mais aussi simplement de le formaliser : pour Ioannis Kanellos, l'intertexte ne peut s'identifier au corpus et aucun macro-corpus ne peut prétendre l'épuiser[23].

---

[21] Si tant est – ce dont doute François Rastier (2001) –, qu'en sémantique, compréhension et interprétation puissent être distinguées. Si le sens n'est pas objectivation mais seulement interprétation, alors le co-texte (environnement textuel nécessaire à l'élaboration et la compréhension d'un texte) n'est plus une notion distincte de l'intertexte (ressources textuelles nécessaires à l'interprétation) mais devient son hyponyme. Dans cet article, nous avons effectivement réfléchi de manière générale à « l'intertexte » pour ne proposer que l'intégration du « co-texte » dans les corpus réflexifs. Le co-texte a été alors compris comme une réduction de l'intertexte à l'environnement linguistique le plus immédiat et le plus objectif d'un texte (par exemple les textes produits par le même auteur, à la même période, dans le même cadre générique ...).

[22] I. Kanellos (1999), « De la vie sociale du texte. L'intertexte comme facteur de la coopération interprétative », *Cahiers de Praxématique*, n°33, p. 58.

[23] *Ibid.*, p. 59.

Pourtant en tant qu'idéal ou ligne d'horizon, la réflexivité du corpus apparaît comme un moyen de faire reculer la subjectivité dans les Sciences du langage ou les Sciences humaines qui manipulent des textes. Elle vise à introduire, dans la sémantique des textes et dans la démarche interprétative, l'objectivité et la transparence qui lui font parfois défaut. L'objectif est en fait d'internaliser autant que possible les ressources sémantiques ou interprétatives cotextuelles en les mettant à disposition du chercheur et de ses lecteurs dans le corpus[24]. *A maxima*, il s'agit là de tirer toutes les conséquences de l'affirmation de F. Rastier : « *Le corpus est la seule objectivation possible (philologique) de l'intertexte, qui sinon demeure une notion des plus vagues* »[25]. *A minima*, il s'agit là d'un simple acte de transparence garantissant une sémantique contrôlée, une interprétation partageable, conditions à une analyse réutilisable.

Les préoccupations actuelles autour du corpus et de l'analyse des textes concernent bien légitimement la nécessaire tentative de contrôle et d'objectivation de l'intertexte. Le corpus réflexif, à son échelle, poursuit cette exigence d'objectivation et de mise en forme des ressources sémantiques et interprétatives, en se proposant d'être, dans la mesure du possible, un tout textuel auto-suffisant – c'est à dire un univers sémantique parmi d'autres – pour une exploitation certes pas exhaustive mais raisonnable et raisonnée du texte.

Cela exige évidemment, au moment de la constitution du corpus un effort de rassemblement de textes tant d'un point de vue qualitatif (quels textes ?) que quantitatif (combien de textes ?). Mais pourquoi serait-il impossible de faire cet effort au départ de l'analyse lorsqu'il paraît naturel de le faire de manière anarchique en cours d'analyse ? Sous la puissance des logiciels de traitement automatique ou semi-automatique du texte et du langage, et l'augmentation des données textuelles effectivement disponibles

---

[24] C'est-à-dire en 1) les rendant consultables et 2) en les traitant scientifiquement par la méthode établie de l'analyse. Dans ce cas de figure, la distinction, selon la terminologie de Rastier & Pincemin (1999), entre *corpus de travail* et *corpus de référence* perdrait de la pertinence.

[25] F. Rastier (1998), « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages*, n°129, p. 17.

sous forme numérisée, la discrimination de traitement entre les données textuelles et les données co-textuelles ou intertextuelles d'une recherche apparaîtra de plus en plus criante et de moins en moins justifiable.

Finalement ramener à sa plus simple expression l'exigence de réflexivité n'est qu'une doléance : celle formulée aux chercheurs de réfléchir au moment de la constitution du corpus aux ressources (co)(inter)(archi)(hyper)textuelles nécessaires à sa compréhension. A partir de là toutes celles qui pourront être explicitées et intégrées au corpus pour subir les mêmes traitements seront les bien venues. Les corpus réflexifs entraîneront peut-être, par le strict contrôle – donc la limitation – des ressources interprétatives, une plus grande modestie au niveau de l'interprétation des textes mais aboutiront sans doute à une herméneutique plus rigoureuse.

Références bibliographiques

Adam J.-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*. Paris : Nathan.

Bilger M. (éd.) (2000). *Corpus. Méthodologie et applications linguistiques*. Paris : Champion.

Cahiers de praxématique (1999). « Sémantique de l'intertexte », n°33.

Condamines A. *et al.* (1999). *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, Actes de l'atelier thématique TALN, Cargèse.

Guilhaumou J., Maldidier D. & Robin R. (1994). *Discours et archive*. Liège : Mardaga.

Habermas J. (1994). *Textes et contextes*. Paris : Seuil.

Habert B. *et al.* (1997). *Les linguistiques de corpus*. Paris : A.Colin.

Heiden S. (1999). « Encodage uniforme et normalisé de corpus. Application à l'étude d'un débat parlementaire ». *Mots* 60 : 113-132.

Kanellos I. (1999). « De la vie sociale du texte. L'intertexte comme facteur de la coopération interprétative ». *Cahiers de Praxématique* 33 : 41-82.

Lahlou S. & Folch H. (1998). « Quelques stratégies pour l'exploitation en ADT de grands corpus hétérogènes ». *In JADT 1998*, Nice : CNRS-UNSA, pp. 381-391.

Mayaffre D. (1999). « Histoire politique ou politique de l'Histoire ? ». *La Pensée* 319 : 139-147.

Mayaffre D. (2000). *Le poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres. Maurice Thorez, Léon Blum, Pierre-Etienne Flandin et André Tardieu (1928-1939).* Paris : Honoré Champion.

Péry-Woodley M.-P. (1995). « Quels corpus pour quels traitements automatiques ? ». *Traitement Automatique des Langues* 36, 1-2 : 213-232.

Peschanski D. (1984). *Histoire moderne et contemporaine et informatique*, n°4 spécial « Le corpus ».

Prost A. (1988). « Les mots ». *In* R. Rémond (sous la direction de), *Pour une histoire politique,* Paris : Seuil.

Pincemin B. (1999). « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative ». *In* A. Condamines *et al., Corpus et traitement..., op. référencé.*

Rastier F. (1998). « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage. ». *Langages* 129 : 97-111.

Rastier F. (2001). *Arts et sciences du texte.* Paris : PUF.

Rastier F. & Pincemin B. (1999). « Des genres à l'intertexte ». *Cahiers de Praxématique* 33 : 83-111.

Robin R. (1973). *Histoire et Linguistique.* Paris : A. Colin.

Tanguy L. &. Thlivitis T. (1999). « Parcours interprétatifs (inter) textuels : vers une assistance informatique ». *Cahiers de praxématique* 33 : 185-215.

Thlivitis T. (1999). *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension de textes*, Thèse de Doctorat Informatique, Univ. De Rennes 1.